Building a database of Alzheimer's patients

Jicheng Lu

1 Introduction

At this stage, we need to build a database of the patients who have diagnosed with Alzheimer's disease. However, due to the privacy issue and limited access to the existing database, we cannot obtain enough data, especially the individual patient's data. We need to find an alternative method that is able to generate an arbitrary number of patient data.

2 Method

The solution introduced here is to generate "fake" data based on the aggregated patient data. In other words, we can sample multiple data points from the group distributions.

For example, suppose we want to have patients with different ages. We first collect the statistical data from relevant medical studies, such as the distribution of age ranges (60% for 80-89, 30% for 70-79, 10% for 60-69). Then we can sample our patient data based on this distribution. One possible result can be 60 patients with age 80-89, 30 patients with age 70-79, 10 patients with age 60-69. We can apply this strategy to other attributes, such as gender, years of education, willingness, MMSE score, etc.

3 Evaluation

Basically, there are two ways of evaluating the rationality of the generated data:

- Subjective evaluation: The data is evaluated by doctors or physicians based on their experience.
- Model evaluation: A classification model is trained based on the generated data and used to evaluate future data.

For the second approach, we can first collect statistical data from Alzheimer's study and non-Alzheimer's study (e.g., other dementia diseases), respectively. Then we can train the model with more reliability.

4 Issues

Here we list several possible issues when sampling data:

- What attributes do we need for each patient? e.g., age, race, gender, MMSE, etc.
- Is there any relationship between attributes? e.g. age vs stroke history, etc.
- How do we connect the generated data with the criteria?

5 Resources

There are two resources I have found. Both of them contain a collection of statistical data from AD-related studies.

- GAAIN: http://www.gaain.org/
- NACC: https://www.alz.washington.edu/WEB/demoDx.html

6 Examples of Data Source

Here we give some examples from the websites mentioned above.

6.1 GAAIN

For GAAIN (i.e, Global Alzheimer's Association Interactive Network), the data is built-in, so we can only check them in their interface. The interface is called "The Interrogator" (register an account to log in). Fig. 1 gives the general picture after we log in. It contains a number of AD-related studies from all over the world.

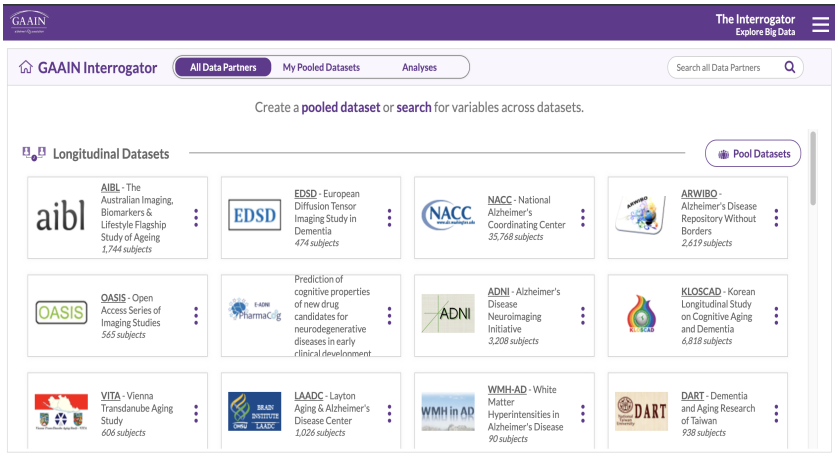

Figure 1: The Interrogator of GAAIN.

After we select a study, we can observe the statistical data it contains. Fig. 2 presents the dataset of AIBL, a medical study from Australia. At the left hand side, we can select different variables and the results are shown at the right side. For example, Fig. 2 shows the gender distribution of the subjects. We can also get the medical data, such as MMSE, medical history, etc.

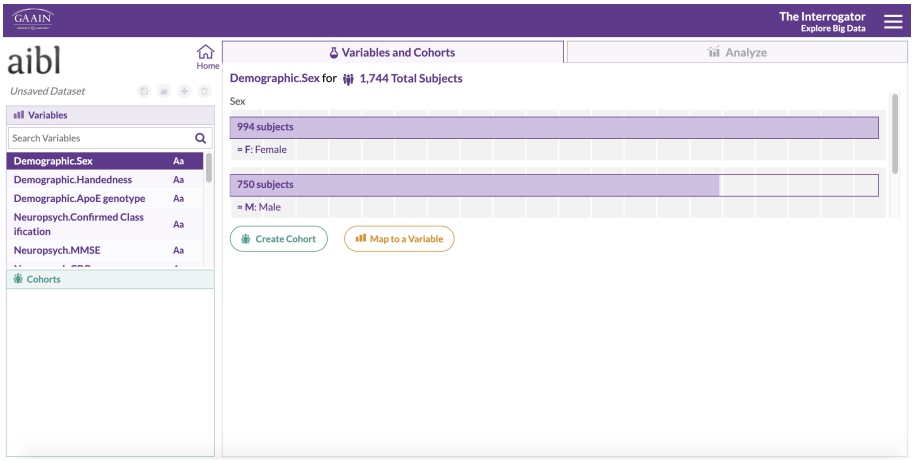

Figure 2: The dataset of AIBL.

6.2 NACC

For NACC (i.e., National Alzheimer's Coordinating Center), they have studied patients with not only the Alzheimer's disease but also the other cognitive illness, such as Parkinson's disease, vascular brain injury, etc. Currently, we can get direct access to the UDS summary data, which contains the subject demographics as well as cognitive status (Fig. 3). The UDS dataset was contributed by the Alzheimer's Disease Centers from 2005 to present. There are also other relevant data on NACC. For more details, please use the link provided above.

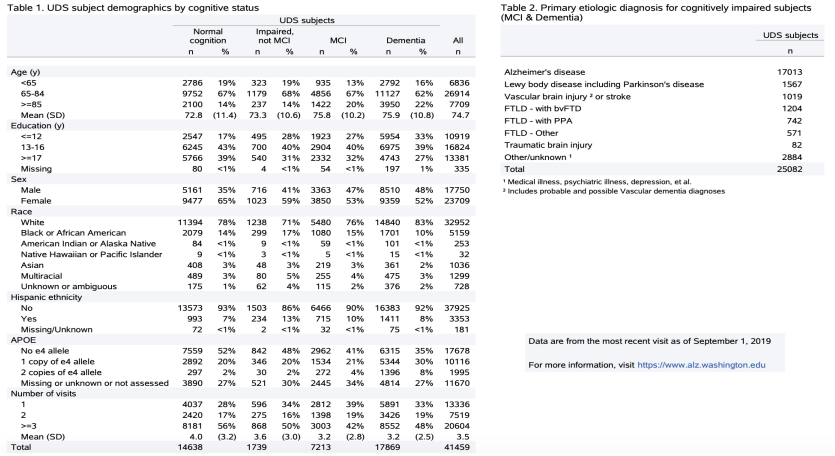

Table 1. UDS subject demographics by cognitive status

	UDS subjects								
	Normal cognition		Impaired, not MCI		MCI		Dementia		All
	n	%	n	%	n	%	n	%	n
Age (y)									
<65	2786	19%	323	19%	935	13%	2792	16%	6836
65-84	9752	67%	1179	68%	4856	67%	11127	62%	26914
>=85	2100	14%	237	14%	1422	20%	3950	22%	7709
Mean (SD)	72.8	(11.4)	73.3	(10.6)	75.8	(10.2)	75.9	(10.8)	74.7
Education (y)									
<=12	2547	17%	495	28%	1923	27%	5954	33%	10919
13-16	6245	43%	700	40%	2904	40%	6975	39%	16824
>=17	5766	39%	540	31%	2332	32%	4743	27%	13381
Missing	80	<1%	4	<1%	54	<1%	197	1%	335
Sex									
Male	5161	35%	716	41%	3363	47%	8510	48%	17750
Female	9477	65%	1023	59%	3850	53%	9359	52%	23709
Race									
White	11394	78%	1238	71%	5480	76%	14840	83%	32952
Black or African American	2079	14%	299	17%	1080	15%	1701	10%	5159
American Indian or Alaska Native	84	<1%	9	<1%	59	<1%	101	<1%	253
Native Hawaiian or Pacific Islander	9	<1%	3	<1%	5	<1%	15	<1%	32
Asian	408	3%	48	3%	219	3%	361	2%	1036
Multiracial	489	3%	80	5%	255	4%	475	3%	1299
Unknown or ambiguous	175	1%	62	4%	115	2%	376	2%	728
Hispanic ethnicity									
No	13573	93%	1503	86%	6466	90%	16383	92%	37925
Yes	993	7%	234	13%	715	10%	1411	8%	3353
Missing/Unknown	72	<1%	2	<1%	32	<1%	75	<1%	181
APOE									
No e4 allele	7559	52%	842	48%	2962	41%	6315	35%	17678
1 copy of e4 allele	2892	20%	346	20%	1534	21%	5344	30%	10116
2 copies of e4 allele	297	2%	30	2%	272	4%	1396	8%	1995
Missing or unknown or not assessed	3890	27%	521	30%	2445	34%	4814	27%	11670
Number of visits									
1	4037	28%	596	34%	2812	39%	5891	33%	13336
2	2420	17%	275	16%	1398	19%	3426	19%	7519
>=3	8181	56%	868	50%	3003	42%	8552	48%	20604
Mean (SD)	4.0	(3.2)	3.6	(3.0)	3.2	(2.8)	3.2	(2.5)	3.5
Total	14638		1739		7213		17869		41459

Table 2. Primary etiologic diagnosis for cognitively impaired subjects (MCI & Dementia)

	UDS subjects
	n
Alzheimer's disease	17013
Lewy body disease including Parkinson's disease	1567
Vascular brain injury [2] or stroke	1019
FTLD - with bvFTD	1204
FTLD - with PPA	742
FTLD - Other	571
Traumatic brain injury	82
Other/unknown [1]	2884
Total	25082

[1] Medical illness, psychiatric illness, depression, et al.
[2] Includes probable and possible Vascular dementia diagnoses

Data are from the most recent visit as of September 1, 2019

For more information, visit https://www.alz.washington.edu

Figure 3: The UDS dataset of NACC.

Moreover, we can request the custom data from NACC. The keywords that can be selected belong to the following categories:

- Clinical Diagnosis, e.g., Alzheimer's disease, Lewy body dementia, etc.
- Clinical Measures and Symptoms, e.g., blood pressure, Hachinski ischemic score, etc.
- Neuropsychological Testing, e.g., MMSE, composite scores, etc.
- Demographics, e.g., sex, race, age, etc.
- Subject health history, e.g., smoking, alcohol, etc.
- Neuropathology / Death, e.g., Lewy bodies, Neuritic plaques, etc.
- Genetics and Biomarkers, e.g., MRI, PET, etc.
- Study Design, e.g., cross-sectional, longitudinal, etc.

The entire keywords are attached with this proposal.

7 Examples of Generated Data

In this section, we present several examples of generated data using the UDS data from NACC (Fig. 3). First, we introduce the method in which we generate the patient data. As we observe from the data, we know that there is no interaction between different attributes (i.e., age, sex, race, ethnicity, years of education, etc.). For an instance, we cannot get one person's age based on his/her race and vice versa. Thus, we treat each attribute independently.

The main idea is to use the Bayes formula (Eq. 1). We want to get the diagnosis from the given attributes. The diagnosis classes are: "normal", "impaired, not MCI", "MCI", "dementia".

$$P(diagnosis|attributes) = \frac{P(diagnosis) \times P(attributes|diagnosis)}{P(attributes)} \tag{1}$$

Since all the attributes are independent, we compute the attribute-related components as follows. We denote the attributes as $a_1, a_2, ..., a_n$.

$$P(attributes) = P(a_1)P(a_2)...P(a_n) \tag{2}$$

$$P(attributes|diagnosis) = P(a_1|diagnosis)P(a_2|diagnosis)...P(a_n|diagnosis) \tag{3}$$

Here we randomly generate 10,000 "fake" patients data, and assign diagnosis to each using Eq. 1. Fig. 4 gives the basic layout of part of the data.

ID	age	yearOfEducation	sex	race	hispanic	diagnosis
1	78	21	Female	Black or African American	No	Normal
2	64	14	Female	Asian	No	Normal
3	76	16	Male	White	No	Dementia
4	90	8	Male	White	No	Dementia
5	94	10	Female	Unknown	No	Dementia
6	75	22	Female	White	No	Normal
7	76	20	Male	White	No	Dementia
8	69	10	Female	White	No	Dementia
9	76	13	Female	White	No	Normal
10	91	14	Male	White	No	Dementia
11	84	19	Female	White	No	Normal
12	84	20	Male	White	No	Dementia
13	76	23	Male	Asian	No	Normal
14	83	20	Male	White	No	Dementia
15	63	20	Female	White	No	Normal
16	76	9	Male	White	No	Dementia
17	68	24	Male	White	Yes	Dementia
18	71	7	Male	White	No	Dementia
19	84	11	Male	White	No	Dementia
20	74	18	Female	White	No	Normal
21	83	14	Female	Black or African American	No	Normal
22	78	24	Male	White	No	Dementia

Figure 4: The layout of the generated data.

To validate the correctness of the generated data, we compare the distributions of several attributes between the generated data and raw data.

Table 1: Distribution comparison of "age".

Age	raw distribution(%)	generated distribution(%)
< 65	16.5	16.2
65-84	64.9	65.2
≥ 85	18.6	18.6

Table 2: Distribution comparison of "years of education".

Education (y)	raw distribution(%)	generated distribution(%)
≤ 12	26.3	26.7
13-16	40.6	40.9
≥ 17	32.3	32.3

Table 3: Distribution comparison of "sex".

Sex	raw distribution(%)	generated distribution(%)
Male	42.8	42.5
Female	57.2	57.5

Moreover, we check the distribution of the generated diagnosis. We notice that there is zero diagnosis of "Impaired, not MCI" and "MCI". This may be because the original probabilities of these two diagnoses are much smaller, compared to the other two cases.

Table 4: Distribution comparison of "diagnosis".

Diagnosis	raw distribution(%)	generated distribution(%)
Normal	35.3	38.2
Impaired, not MCI	4.2	0.0
MCI	17.4	0.0
Dementia	43.1	61.8

8 Summary

This proposal introduces the idea about generating "fake" patients data from the aggregated data. The data is generated using Bayes formula, where each attribute is treated independently. We provide several examples of the generated data, and we observe that they follow the similar distributions with the raw data.